TEACHING LOGISTIC REGRESSION

Michael J. Campbell, Northern General Hospital, UK

Based on the experience of teaching logistic regression to non-mathematicians, a number of areas of possible confusion are identified that may arise particularly when the method is contrasted with multiple linear regression. The fact that the model is multiplicative in odds ratios means that the concept of interaction needs to be clearly defined. Confidence intervals for the estimates of the odds ratios are asymmetric about the estimate, in contrast to confidence intervals in multiple regression which are symmetric. The fact that including a covariate will often increase the standard error of an estimate, rather than decrease it, is somewhat counter-intuitive. Logistic regression must be clearly distinguished from logit, or log-linear modelling.

INTRODUCTION

I have taught logistic regression on a number of postgraduate courses, to doctors and health care professionals on Masters in Public Health and in Epidemiology. I have discovered a number of areas in which confusion reigns, and in this paper I intend to highlight problematic areas.

THE MODEL

In all regression models we distinguish between a *dependent* or outcome variable, and an *independent* or input variables. In multiple linear regression the dependent variable is continuous, and the independent variables can nominal, categorical or continuous. In logistic regression the independent variable is nominal, that is it can take one of two categories. As an example Oakeshott et al (1998) looked at various risk factors for chlamydial infection in a cross-sectional survey. The dependent variable was the presence or absence of *Chlamydia trachomatis* on a cervical smear. The predictor variables were age<25, race and number of sexual partners. The focus is on the risk of Chlamydia infection, given age, for example. Each of these associations was tested separately using a chi-squared test and found to be significant. The questions that remained included:
i) Are any of the variables confounded (for example age and race), so that if we control for age, is race still significant? ii) Is there any interaction between the input variables, for example is a young person with multiple partners at much higher risk than would be predicted from each risk factor separately?

When the input variables are categorical, one can cross-tabulate them with the outcome variable, and determine, for example, the proportion of women with Chlamydia, who are aged less than 25, of one particular and with more than 2 sexual partners.

The model needs to be described with care. If the expected (or population) probability of a positive result for a woman i with risk factors $X_{i1}...,X_{ip}$ is π_i then the model is

$$\text{logit}(\pi_i) = \log_e\{\pi_i/(1-\pi_i)\} = \beta_0 + \beta_1 X_{i1}+...+\beta_{ip}X_{ip}. \quad (1)$$

I initially justify the logit transform by stating that the right hand side of equation (1) is unbounded. A probability must lie between 0 and 1. An odds ratio must lie between 0 and infinity. A log odds ratio, or logit, is unbounded and has the same range as the right hand side of equation (1).

Most elementary books distinguish the case when all the independent variables are categorical and when some are continuous. If the independent variables are categorical, and the data tabulated by all levels of the covariables, then all women in a cell of the table will have the same value of π_i and this can be estimated by p_i which is the proportion of women who have Chlamydia in that cell. If the independent variables are continuous, then such a table cannot be drawn up; if one attempted to do so there would be as many cells as there were women (if there were no ties in the continuous variable) and so the proportions would be zero or one. Often, by analogy to multiple regression, the model is described in the literature as above, but with the observed proportion, p_i, replacing p_i. This is incorrect, because a model gives a prediction, and the relationship between the prediction and the observed dependent variable depends on the error distribution. Many books introduce logistic regression immediately after linear regression, and appear to suggest that the two are very similar. However the analogy is not exact and this can cause confusion amongst students. It is not so easy as in multiple regression to add an ‘error’ term’, and so the concept of minimising the error is more difficult. In theory the parameters are estimated by maximising the likelihood of the observed values, using a binomial error distribution. Students may not have studied maximum likelihood at this stage, (or indeed may never study it) and so the estimation procedure may appear mysterious. It can be helpful to demonstrate that the weighted least squares approach does, in fact give sensible results. Another problem is when the dependent variable is 0/1, the student discovers that the logit of the dependent variable does not exist. This may lead them to believe that logistic regression is impossible in these circumstances. Defining

residuals is more difficult in logistic regression and model checking is different to the linear regression situation.

LOGIT/LOGISTIC

Some computer programs confuse logit (or log-linear) models and logistic regression. Logit models are used to analyse large contingency tables. They differ from logistic models in that;

1. There is no clear division between dependent and independent variables.
2. In Logit models one has to fit the marginal values first, and associations are measured by interactions. Thus for a logit model, in the Chlamydia example, one would have to fit parameters corresponding to the proportion of subjects with Chlamydia and aged <25 before fitting a parameter corresponding to the interaction between the two. In logistic regression, the presence or absence of Chlamydia is unequivocally the dependent variable, and marginal values do not have to be accounted for.
3. In logistic regression the independent variables can be continuous.

CONSEQUENCES OF THE LOGISTIC MODEL

It is a standard result that if b_1 is the estimate of β_1 then $\exp(b_1)$ is the estimated odds ratio associated with X_1 (note *not* relative risk as is sometimes stated). If X_1 is continuous, then it is the increase in odds associated with a unit increase in X_1. It is also a standard result to show that, in a case control study, if the dependent variable is case/control status, then $\exp(b_1)$ again gives a valid estimate of the odds ratio associated with X_1. For service courses I simply state these results and do not attempt to prove them! Since the model is described in terms of logs, what is additive on log scale is multiplicative on the linear scale. Thus if being in a particular racial group increases the risk of Chlamydia by 2 and being aged less than 25 increases the risk by a factor of 3, then if the two risk factors are independent, someone in a particular racial group, and being aged less than 25 will have a risk 2x3=6 times that of someone without those risk factors. It is important to stress to students that this is *not* an interaction, this is a consequence of the model if the two factors are independent. If RACE takes the value 1 for someone who is of a particular race and zero elsewhere, and AGE takes the value 1 for someone aged <25 and 0 elsewhere, then a model with AGE and RACE implies multiplicative risks. The

interactive term AGExRACE can be included to see whether there is synergy (more than multiplicative) or whether the two factors interact together to be less than multiplicative. Another consequence of the model is that the output is usually of the form *b* and SE(*b*). A 95% confidence interval for *b* is given by b+/-1.96xSE(*b*). Thus a 95% confidence interval for the Odds Ratio is exp{*b*-1.96xSE*b*)} to exp{*b*+1.96xSE(*b*)}. This is asymmetric about OR, which can cause some confusion amongst the students. For example, from Oakeshott(1998), the Odds Ratio for Chlamydia for someone aged<25 is 2.5, 95% CI 1.1 to 5.6.

MODEL CHECKING

An important question is whether the model describes the data well. This has been extensively described by Collett(1991). If the logistic model is grouped, then there is no problem comparing the observed proportions in the groups and those predicted by the model. However, if some of the input variables are continuous, one has to group the predicted values in some way. Hosmer and Lemeshow(1989) suggest a number of methods. In practice, investigators use the Hosmer-Lemeshow statistic to reassure themselves that the model describes the data and so they can interpret the coefficients. However, there is a theoretical objection to using a significance test to determine goodness of fit, before using another test to determine whether coefficients are significant. If the first test is not significant, it does not tell us that the model is true, only that we do not have enough evidence to reject it. Since no model is exactly true, with enough data the goodness of fit test will always reject the model, but the model may be 'good enough' for a valid analysis. Also, if the model does not fit, what do we do? Is it invalid to make inferences from the model? Collett (1991) describes a number of ways to investigate when the model departs from the data more carefully and ways of correcting for departures.

There are a number of ways the model may fail to describe the data well: i) lack of an important covariate, ii) outlying observations, iii) 'extra-binomial' variation (Williams 1982), iv) the logistic transform is inappropriate. The first problem can be investigated by trying all available covariates, and interactions between them. Provided the omitted covariate is not a confounder, then inference about the covariate of interest is usually not affected. For example if the proportion of people aged <25 in the study by Oakeshott et al (1998) was the same in each racial group, (that is if a subject were in the

survey and aged <25, they would not be more likely to be in one particular racial group than any other) then the estimated the risk of chlamydial infection for people aged <25 will not be affected by whether race is included in the model.

Outlying observations can be difficult to check when the outcome variable is 0/1. However, some packages do provide standardised residuals and these can be plotted. It is important also to look for influential observations which if deleted would change the parameter estimates. Details are given by Pregibon(1981).

Extra-binomial variation can occur when the data are not strictly independent. For example, repeated outcomes within an individual, or patients grouped by general practitioner. Whilst the estimate of the regression coefficients is not unduly affected, the estimates of the standard errors are usually underestimated, leading to a type I error rate higher than the expected 5% (Cox and Snell, 1989) In the past this has been dealt with by an approximate method, for example by scaling the standard errors (Williams, 1982). However, it is now viewed as a special case of what is known as a *random effects* model in which one (or more) of the regression coefficients *b* is regarded as random with a mean and variance that can be estimated, rather than fixed. Many statistical packages have yet to accommodate this type of model in logistic regression, although ones that do include STATA5 and EGRET .

COVARIATE ADJUSTMENT

In multiple regression it is natural to adjust the effect of one covariate for the influence of another. However, this is not necessarily the case in logistic regression (Robinson and Jewell 1991). Let Y, X_1 and X_2 each be a dichotomous variable taking the value 0 or 1. The variable Y is the outcome variable and X_1 the risk factor of principal interest.

Suppose the following two models both provide a valid description of the population structure.

$$\text{logit}(\pi)=\beta^0_0 + \beta^0_1 X_1 \qquad (2)$$

$$\text{logit}(\pi)=\beta^1_0 + \beta^1_1 X_1 + \beta^1_2 X_2 \; . \qquad (3)$$

If the estimated values of the coefficient associated with X_1 in (2) and (3) are b^0_1 and b^1_1, then Robinson and Jewell (1991) showed that $\text{Var}(b^0_1) <= \text{Var}(b^1_1)$, in other words allowing for a covariate will almost always *increase* the standard error of the estimate of interest. However, this does not mean to say that one should not adjust for

covariates. They point out that, if X_1 and X_2 are independent, then b^0_1 will fall between b^1_1 and zero. Thus although the variance of b^0_1 is smaller, so is its estimate, and so a significance test of an effect without the covariate in the model may given a large p-value than one with a covariate in the model. The conflict is between bias and precision. The estimated parameter b^1_1 is unbiased but less precise than the biased estimate b^0_1. If the study is large, then bias is usually more important than precision, and so Robinson and Jewell(1991) conclude that in general it is still a good idea to include covariates!

DISCUSSION

Before logistic regression is taught, the students should be familiar with the following concepts: logarithms, odds, the binomial distribution, multiple linear regression and maximum likelihood. For teaching on service courses the latter requirement may be omitted but then it is easiest to restrict attention to the situation where there are a limited number of independent variables, none of which are continuous. In this case the data can be tabulated, the links with the chi-squared test can be made, and model checking is easier.

REFERENCES

Clayton, D. and Hills M. (1993). *Statistical Models in Epidemiology*. Oxford: Oxford Science Publications .

Collett, D. (1991). *Modelling Binary Data*. London: Chapman and Hall.

Cox, D. R. and Snell E. J.(1989). *Analysis of Binary Data* (2nd Ed) London: Chapman and Hall .

Hosmer, D. W. and Lemeshow, S. (1989). *Applied Logistic Regression* New York: John Wiley and Sons.

Oakeshott, P., Kerry S, Hay S., and Hay P. (1998). Opportunistic screening for chlamydial infection at time of cervical smear testing in general practice: prevalence study. *British Medical Journal, 316,* 351-352.

Pregibon, D. (1981). Logistic regression diagnostics. *Annals of Statistics, 9,* 705-724.

Robinson, L. D. and Jewell, N. P. (1991). Some surprising results about covariance adjustment in logistic regression models. *International Statistical Review, 59,* 227-240.

Williams, D. A.(1982). Extra-binomial variation in logistic linear models. *Applied Statistics, 31,* 144-8.